Robust Classification of High Dimensional Unbalanced Single and Multi-label Datasets

A Thesis Submitted for the Degree of
Doctor of Philosophy

By
Ali Braytee

in

School of Software
UNIVERSITY OF TECHNOLOGY SYDNEY
AUSTRALIA
FEBRUARY 2018

UNIVERSITY OF TECHNOLOGY SYDNEY
SCHOOL OF SOFTWARE

The undersigned hereby certify that they have read this thesis entitled "**Robust Classification of High Dimensional Unbalanced Single and Multi-label Datasets**" by **Ali Braytee** and that in their opinions it is fully adequate, in scope and in quality, as a thesis for the degree of **Doctor of Philosophy**.

Dated: February 2018

Principal Supervisor: ______________________________
A/P Paul Kennedy

CERTIFICATE

Date: **February 2018**

Author: **Ali Braytee**

Title: **Robust Classification of High Dimensional Unbalanced Single and Multi-label Datasets**

Degree: **Ph.D.**

I, Ali Braytee declare that this thesis, submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Software/ Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by an Australian Government Research Training Program Scholarship.

Signature of Author

Acknowledgements

My first acknowledgement is to God for being my source of spiritual strength. I am greatly indebted to my supervisor, A/P Paul Kennedy for his continuous encouragement, advice, help and invaluable suggestions. I owe my research achievements to his experienced supervision. Many thanks are also due to my co-supervisor, Dr. Wei Liu for his valued suggestions and constant support, and for the numerous conversations with him. I gratefully acknowledge the useful discussions with A/P Daniel Catchpoole and A/P Farookh Hussain. I appreciate the travel support for attending the international conferences which I received from the School of Software and the Centre for Artificial Intelligence.

I would like to thank my parents that I could never have done this without their faith, support, and constant prays. I also thank my wife Zaynab, my partner in life, who managed to be nothing but supportive, for your patience, and your faith, because you always understood.

Last but not least, the financial assistance of an Australian Postgraduate Award and UTS Research Excellence Scholarship is gratefully appreciated.

To my father and mother who instilled in me the virtues of perseverance and commitment. To my wife Zaynab, whose sacrificial care for me and our children enabled the hours of research, contemplation, and writing necessary to complete this thesis. To my children Hassan and Adam you are my greatest achievement, and you will always be my blessings from God. All together we have managed to survive throughout this challenging journey. I love you all.

Table of Contents

List of Tables

List of Figures

Abstract

Single and multi-label classification are arguably two of the most important topics within the field of machine learning. Single-label classification refers to the case where each sample is assigned to one class, and multi-label classification is where instances are associated with multiple labels simultaneously. Nowadays, research to build robust single and multi-label classification models is still ongoing in the data analytics community because of the emerging complexities in the real-world data, and due to the increasingly research interest in use of data analytics techniques in many fields including biomedicine, finance, text mining, text categorization, and images. Real-world datasets contain complexities which degrade the performance of classifiers. These complexities or open challenges are: imbalanced data, low numbers of samples, high-dimensionality, highly correlated features, label correlations, and missing labels in multi-label space. Several research gaps are identified and motivate this thesis. Class imbalance occurs when the distribution of classes is not uniform among samples. Feature extraction is used to reduce the dimensionality of data. However, the presence of highly imbalanced data in single-label classification misleads existing unsupervised and supervised feature extraction techniques. It produces features biased towards classification of the class with the majority of samples, and results

in poor classification performance especially for the minor class. Furthermore, imbalanced multi-labeled data is more ubiquitous than single-labeled data because of several issues including label correlation, incomplete multi-label matrices, and noisy and irrelevant features.

High-dimensional highly correlated data exist in several domains such as genomics. Many feature selection techniques consider correlated features as redundant and therefore need to be removed. Several studies investigate the interpretation of the correlated features in domains such as genomics, but investigating the classification capabilities of the correlated feature groups in single-labeled data is a point of interest in several domains. Moreover, high-dimensional multi-labeled data is more challenging than single-labeled data. Only relatively few feature selection methods have been proposed to select the discriminative features among multiple labels due to issues including interdependent labels, different instances sharing different label correlations, correlated features, and missing and noisy labels.

This thesis proposes a series of novel algorithms for machine learning to handle the negative effects of the above mentioned problems and improves the performance of the classifiers in single and multi-labeled data. There are seven contributions in this thesis. Contribution 1 proposes novel cost-sensitive principal component analysis (CSPCA) and cost-sensitive non-negative matrix factorization (CSNMF) methods for handling feature extraction of imbalanced single-labeled data. Contribution 2 extends a standard non-negative matrix factorization to a balanced supervised non-negative matrix factorization (BSNMF) to handle the class imbalance problem in supervised non-negative matrix factorization. Contribution 3 introduces an ABC-Sampling algorithm for balancing imbalanced datasets based on Artificial Bee Colony algorithm.

Contribution 4 develops a novel supervised feature selection algorithm (SCANMF) by jointly integrating correlation network and structural analysis of the balanced supervised non-negative matrix factorization to handle high-dimensional, highly correlated single-labeled data. Contribution 5 proposes an ensemble feature ranking method using co-expression networks to select optimal features for classification. Contribution 6 proposes a Correlated- and Multi-label Feature Selection method (CMFS), based on NMF for simultaneously performing multi-label feature selection and addressing the following challenges: interdependent labels, different instances sharing different label correlations, correlated features, and missing and flawed labels. Contribution 7 presents an integrated multi-label approach (ML-CIB) for simultaneously training the multi-label classification model and addressing the following challenges namely, class imbalance, label correlation, incomplete multi-label matrices, and noisy and irrelevant features.

The performance of all novel algorithms in this thesis is evaluated in terms of single and multi-label classification accuracy. The proposed algorithms are evaluated in the context of a childhood leukaemia dataset from The Children Hospital at Westmead, and public datasets for different fields including genomics, finance, text mining, images, and others from online repositories. Moreover, all the results of the proposed algorithms in this thesis are compared to state-of-the-art methods. The experimental results indicate that the proposed algorithms outperform the state-of-the-art methods. Further, several statistical tests including, t-test and Friedman test are applied to evaluate the results to demonstrate the statistical significance of the proposed methods in this thesis.

Publications

Below is the list of journal and conference papers associated with my PhD research:

1. **Braytee, A.**, Liu, W., Anaissi, A. & Kennedy, P. J. (2017), 'Correlated Multi-label Classification with Missing Labels and Class Imbalance', *Data Mining and Knowledge Discovery*. (Under review).

2. **Braytee, A.**, Liu, W., Catchpoole, D. R. & Kennedy, P. J. (2017), Multi-Label Feature Selection using Correlation Information, in 'Proceedings of the 26th ACM International on Conference on Information and Knowledge Management', ACM, (To appear). (ERA rank A).

3. **Braytee, A.**, Liu, W. & Kennedy, P. J. (2017), Supervised context-aware non-negative matrix factorization to handle high-dimensional high-correlated imbalanced biomedical data, in 'International Joint Conference on Neural Networks (IJCNN), 2017', IEEE, pp. 4512−4519. (ERA rank A).

4. **Braytee, A.**, Catchpoole, D. R., Kennedy, P. J. & Liu, W. (2016), Balanced supervised non-negative matrix factorization for childhood leukaemia patients, in 'Proceedings of the 25th ACM International on Conference on Information and Knowledge Management', ACM, pp. 2405−2408. (ERA rank A).

5. **Braytee, A.**, Liu, W. & Kennedy, P. (2016), A cost-sensitive learning strategy for feature extraction from imbalanced data, in 'International Conference on Neural Information Processing', Springer, pp. 78−86. (ERA rank A).

6. Anaissi, A., Goyal, M., Catchpoole, D. R., **Braytee, A.** & Kennedy, P. J. (2016), 'Ensemble feature learning of genomic data using support vector machine', *PloS one* **11**(6), e0157330.

7. **Braytee, A.**, Hussain, F. K., Anaissi, A. & Kennedy, P. J. (2015), ABC-sampling for balancing imbalanced datasets based on artificial bee colony algorithm, in 'IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015,IEEE, pp. 594−599.

Table of Symbols

Symbols	Description
X	Data matrix
n, N	Samples or instances in the matrix
m, M	Features or dimensions in the matrix
p, k, K, R	approximate rank for factorization
U	Factorized matrix
V	Factorized matrix
r	Uniform random number between [0,1]
D	Problem dimensionality
FS	Food source (possible solution)
ABC	Artificial Bee Colony
NMF	Non-negative matrix factorization
PCA	Principle component analysis
GA	Genetic algorithm
PSO	Particle swarm optimization
ACO	Ant Colony Optimization
DE	Differential algorithm
SVM	Support vector machine
y	Class vector
α	Majority classes weight
C^+	Positive imbalance cost ratio
C^-	Negative imbalance cost ratio

W	Factorized matrix
H	Factorized matrix
X'	Weighted data matrix
ADASYN	Adaptive Synthetic Sampling Approach for Imbalanced Learning
CCPDT	Class Confidence Proportion Decision Tree
RU	Random undersampling
ML-CST	Maximum likelihood in cost-sensitive learning
Y, Z	Data matrices
A	Common factorized matrix between Y and Z
B, C	Factorized matrices
X^+	data matrix of positive samples
X^-	data matrix of negative samples
σ^+	weight for positive samples
σ^-	weight for negative samples
$LIMIT$	the number of times of using the possible solution
$Abandoned$	the existing solution that reach to maximum $LIMIT$
$FoodSourceSize$	the size of possible solutions
$Best_FS$	A solution with best fitness value
BIRF	A Balanced Iterative Random Forest
T	Arbitrary data matrix
D	Diagonal matrix
Tr	Trace operation of matrix
$corr$	Correlation function such as Pearson correlation
θ	Power of correlation coefficient
A	Adjacency matrix
n^-,N^-	is the number of negative samples
n^+,N^+	is the number of positive samples
$l_{2,1}$	norm regularization
l_1	norm regularization

β	control the contribution of $l_{2,1}$ norm
α	control the contribution of the network structure
AUC	Area under curve
WGCNA	Weighted correlation network analysis
HLR L1/2 L2	Combination of $l_{2,1}$ and l_2 norm regularization
SVD	Singular value decomposition
SVM-RFE	Feature ranking based on SVM
L, P	A matrix to absorb the different scales of matrices
Q	Feature combinations matrix
c	Number of clusters
Y	Multi-label matrix (Chapter 5)
B	Clusters of labels matrix (Chapter 5)
R	Graph Laplacian (Chapter 5)
S	Similarity matrix (Chapter 5)
G	Diagonal matrix (Chapter 5)
α	Control the contribution of label correlation (Chapter 5)
ϵ	Control the contribution of the network structure (Chapter 5)
γ	Control the contribution of sparseness of the model (Chapter 5)
$\hat{Y}$	New multi-label matrix
L	Similarity matrix (Chapter 5)
W	Label-specific features (Chapter 5)
V	Label regularization (Chapter 5)
α	Control the contribution of the new label matrix manifold (Chapter 5)
β	Control the contribution of the difference between the new and original label matrix (Chapter 5)
MLSMOTE	Multi-label SMOTE